Lucia Ku
Professor Avinash Jairam
CIS 3120
16 April 2022

Homework 2 - Web Scraping

For the first part of Homework 2, I chose to use a Wikipedia page (COVID-19 Pandemic by Country and Territory) to scrape data off of. I specifically chose the table "2021 (second-half) monthly cumulative COVID-19 deaths," which shows the amount of deaths from COVID-19 in each country by month, because it was a complete table that can give an accurate idea of how much the pandemic has impacted the entire world. By using this data, individuals would gain an accurate understanding of the impact of this virus.

In the first block of code, I imported the libraries needed to run my code: BeautifulSoup, requests, pandas, and files. In the second block of code, I set the name of the URL of the Wikipedia table that I chose as "covid." In the third block of code, I first defined "scrape" as a way to scrape the URL of the Wikipedia table and underneath it, I then created empty lists for each country and its months. Therefore, seven empty lists were made: "location," "july," "aug," "sept," "oct," "nov," and "dec." Next, I used the requests library to get information from the webpage and used BeautifulSoup to parse through the webpage's data. After that, I used BeautifulSoup to locate the specific table that I wanted to scrape, which was listed under "2021. 2nd half." Then, I used a for loop in order to go through the content of the table, and then an inner for loop to go through the rows of the table in order to gather data from the table. I saved the data into an array. I then used the if function to exclude the first two rows of the Wikipedia table, which are called "Location" and "World," since those pieces of information were irrelevant to the information I was trying to scrape. Inside the if function, I used another for loop and if function in order to remove the commas from the numbers in the dataset, so as to make it easier for me to later conduct calculations on these numbers via Python. After that if function, I created another if function in order to address countries that contain more than one word in their name. If a country has a length that is greater than seven (including the months), then the code will join the names with multiple words into one singular name. Otherwise, every extra word in every country's name would have become an additional column. I also appended the country names and six months into each of their respective lists and converted the month values into floats in order to make it easier for me to conduct calculations later. Next, I exited out of the entire outer for loop and named my csv file "covid19.csv." Then, I created an empty dataframe called "corona_df" and added the previously filled lists from the beginning of the block into the dataframe. After that, I printed the dataframe (which will give me a dataframe output), converted the dataframe into a csv file, and then downloaded it. Finally, I used pandas to make the remaining calculations from the data I just scraped. Through this method, I was able to find and print November and December's count, mean, standard deviation, minimum, maximum, 25th percentile, 50th percentile, and 75th percentile.

By creating this program that scrapes information about the number of deaths from COVID-19 the world has experienced last year, individuals can use this information to discern if there are any trends to take note in regards to the pandemic. Are the rate of deaths slowing? Does this information mean that the vaccines are working and the world is starting to move past the pandemic? It can also serve as a reminder to still remain careful despite being 2-3 years into the pandemic. Hopefully, after seeing this data, individuals will strive to become more vigilant in preventing the further spread of COVID-19.